Journal of Forestry Research (2012) 23(2): 191−196
DOI 10.1007/s11676-012-0232-x

ORIGINAL PAPER

Estimating canopy closure density and above-ground tree biomass using partial least square methods in Chinese boreal forests

LEI Cheng-liang • JU Cun-yong • CAI Ti-jiu • JING Xia
WEI Xiao-hua • DI Xue-ying

Received: 2011-09-19; Accepted: 2011-11-12

Abstract: Boreal forests play an important role in global environment systems. Understanding boreal forest ecosystem structure and function requires accurate monitoring and estimating of forest canopy and biomass. We used partial least square regression (PLSR) models to relate forest parameters, i.e. canopy closure density and above ground tree biomass, to Landsat ETM+ data. The established models were optimized according to the variable importance for projection (VIP) criterion and the bootstrap method, and their performance was compared using several statistical indices. All variables selected by the VIP criterion passed the bootstrap test (p<0.05). The simplified models without insignificant variables (VIP <1) performed as well as the full model but with less computation time. The relative root mean square error (RMSE%) was 29% for canopy closure density, and 58% for above-ground tree biomass. We conclude that PLSR can be an effective method for estimating canopy closure density and above-ground biomass.

Keywords: above-ground tree biomass; bootstrap method; canopy closure density; partial least square regression (PLSR); VIP criterion

Foundation project: This study was supported by the 948 Program of the State Forestry Administration (2009-4-43) and the National Natural Science Foundation of China (No. 30870420).

The online version is available at http://www.springerlink.com

LEI Cheng-liang • JU Cun-yong (✉) • CAI Ti-jiu • DI Xue-ying
School of Forestry, Northeast Forestry University, Harbin 150040, P.R. China. Email: jucunyong@126.com

JING Xia
College of Geomatics, Xi'an University of Science and Technology, Xi'an 710054, P.R. China

WEI Xiao-hua
Department of Earth and Environmental Science, University of British Columbia (Okanagan), 3333 University way, Kelowna, British Columbia, Canada, V1V 1V7.

Responsible editor: Hu Yanbo

Introduction

Forest parameters such as total wood volume, canopy closure density and above-ground tree biomass are often used for forest management and planning, and ecologically related studies (Jakubauskas and Price 1997; Reese et al. 2002; LeMay et al. 2008). Field measurements of forest parameters are costly and time consuming. In recent years, some studies began to investigate forest parameters with satellite data at pixel, stand, and landscape levels (Zhao and Li. 2001; Næsset et al. 2005; Carreiras et al. 2006; LeMay et al. 2008). However, these previous studies focused on single-species forests, while little attention was drawn to conifer and broadleaf mixed forests in the natural boreal forest ecosystem.

Multiple alternative empirical methods have been proposed to deduce the relationships between remotely sensed data and forest parameters derived from inventory data (Jakubauskas and Price 1997; Næsset et al. 2005). Joshi et al. (2006) evaluated the performance of four alternative methods, i.e. neural network, multiple linear regression, forest canopy closure density mapper and maximum likelihood classification on estimation of forest canopy closure density, and concluded that the neural network had the highest accuracy but all needed relatively large datasets. To date, because of the complexity and the uncertainty of remote sensing data, there has been no consensus on the best method for estimating forest parameters. Thus, it is necessary to develop and compare modeling methods and select the best one for wider application of forest parameter estimation.

Several studies verified that Partial Least Squares (PLS) is one of the most efficient methods for extracting and creating reliable models in a wide range of research fields (Hansen and Schjoerring 2003; Næsset et al. 2005; Farifteh et al. 2007). The objectives of our study were: (1) to evaluate the utility of PLS for describing relations between remote sensing data and forest inventory data; (2) to compare two methods using several selected optimal variables; and (3) to simulate the distribution of canopy closure density and above-ground biomass of the study area, which was located in the Daxing'an mountains of northeast China.

Materials and methods

Study area

The study area, as part of the Daxing'an mountain range, was located at Tahe County (52°09′ to 53°23′ N and 123°20′ to 125°07′E), in northwest Heilongjiang Province, China (Fig. 1). The measured study area is 14,420 km^2. The climate is cold temperate, with mean monthly temperature of -24°C for January and 20°C for July, and mean annual precipitation of 428 mm (Yu et al. 2007). The terrain is gentle foothills with elevations ranging from 450 m to 1,000 m above sea level and slope density less than 15° in most places. The dominant tree species are Chinese larch (*Larix gmelinii*), Mongolian pine (*Pinus sylvestris* var. Mongolica), birch (*Betula platyphylla*), and aspen (*Populus davidiana*).

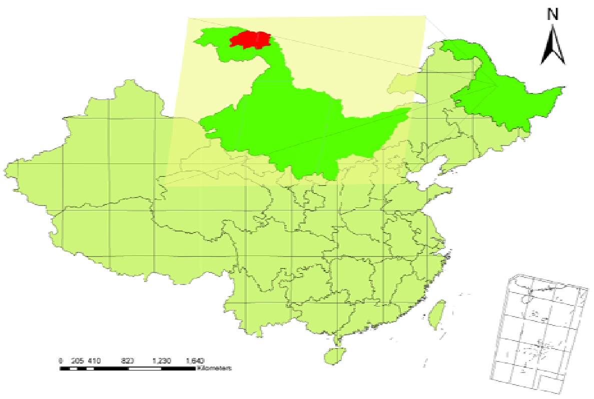

Fig. 1 Location of the study area at Tahe county (red area) Northeast of China

Forest field inventory and Lansat-7 ETM+ data

Data from 263 sample plots in the study area were collected from the forest inventory conducted by the Tahe Forest Service in 2002. These permanent sample plots were spread throughout the study area, and they were included in the relative subcompartments that were fundamental units for management and distinguished by comparability or coherence in many forest characteristics such as species, age grade, and stand site condition (Wu 1996). The sampling plots located within subcompartments represented the variety of terrain and growing conditions in the study area. In each sampling plot we collected data: plot center coordinates, tree species, canopy closure density (the percentage of horizontal surface area covered by tree crowns), wood volume, and terrain (e.g., elevation, slope density, slope position and aspect). Trees of diameter less than 4 cm at breast height (1.3 m above-ground) were not measured, thus they were not taken into account in calculations of wood volume. Above-ground tree biomass (AGB) in t·ha^{-1} was derived from wood volume using documented equations (Fang et al. 1996). The size of each plot was 28.26 m ×28.26 m , i.e. ~ 0.08 ha. However, because each subcompartment had similar tree species composition, age class and site quality class, the potential location deviation of plots inside subcompartments did not cause the problem of spatial matching to remote sensing data.

Two nearly cloud-free Lansat-7 ETM+ scenes were obtained for June 20, 2002, path 121, and rows 23 and 24. The following bands were used in the analysis: TM1 (0.45–0.515 μm), TM2 (0.525–0.605 μm), TM3 (0.63–0.69 μm), TM4 (0.775–0.90 μm), TM5 (1.55–1.75 μm), TM7 (2.09–2.35 μm). The two scenes were geometrically registered respectively by two order polynomial methods referring to distinct ground control points whose coordinates were measured on a 1:50000 scale terrain map, and the resampling method was nearest neighbor. Their geometric accuracy is about 17 m, near half a pixel. Then both the scenes were mosaiced and the digital numbers of overlapping pixels were defined as average values of the two relevant values from different images. Other than geometric registration, no corrections were applied to the remote sensing images.

Digital numbers (DNs) values corresponding to 263 plots for each band were extracted from the ETM+ data, and several spectral indices were computed using the methods described in Table 1. Spectral indices were combined with several terrain characteristic variables, and were then used to deduce the relationships between forest parameters (i.e. canopy closure density and AGB) and remotely sensed data using the partial least square regression method. All the explanatory variables and their derivation are listed in Table 1.

Table 1. Summary of explanatory variables

Variables	Simple description or formula	References
TM1,TM2,…,TM5,TM7	Digital number of ETM+ bands, except band 6	
KT1,KT2,…,KT6	Tassel Cap Transformation of the above six bands	Crist and Cicone 1984; Cohen et al., 2003
TM4-3/4+3	(TM4-TM3)/(TM4+TM3)	Carreiras et al., 2006;
TM4/2	TM4/TM2	Price et al., 2002
TM4/5	TM4/TM5	Price et al., 2002
TM231	(TM2-TM3)/(TM2-TM3+TM1)	LeMay et al., 2008
TM57	(TM5-TM7)/(TM5+TM7)	Hernandez-Stefanoni and Ponce-Hernandez, 2006
TM47	(TM4-TM7)/(TM4+TM7)	Carreiras et al., 2006; LeMay et al., 2008
TM4*3/7	TM4*TM3/TM7	Zhao and Li, 2001
ELEVATION	Elevation above sea level	
SLOPE	Slope density	
UPPER, MIDDLE,LOWER	The upper, middle, lower of slope position	

PLS regression and selection of variables

Collinearity among derived spectral explanatory variables can invalidate ordinary least square estimates, but PLS regression can avoid this problem by compressing these collinear variables into a few orthogonal and non-correlated latent variables (Næsset et al. 2005; Cho et al. 2007). Hence, a PLS model was used to simulate the relationships between ground inventory data and remote sensing data. For model simplification, the bootstrap method was used to test the significance of explanatory variables and to identify dependent variables because it does not specify data distribution conditions (Serneels and Van Espen 2005; Wang et al. 2006). Simultaneously, the VIP (Variable Importance for the Projection) criterion was also considered. Variables of VIP<1 were excluded from the model because they made insignificant or uncertain contribution to the response variables (SIMCA-P 11.5, Umetrics Inc., 2007). Correspondingly, the model containing all variables was named as a full model, and the model without insignificant variables was named as a reduced model. Furthermore, the reduced models were named as the reduced bootstrap model and the reduced VIP model in terms of the criterion for selecting variables, respectively.

Validation and application to the study area

Data collected at the 263 plots were split into calibration data (171, about 65% of all) and validation test data (35%). Calibration data were used to deduce the relationships between explanatory and response variables, while validation data were used to assess agreement between the measured and predicted values from the models. Differences between measured and predicted values were evaluated by average deviation (AD), maximum absolute deviation (MAD), average absolute deviation (AAD), and root mean square error (RMSE). These validation indices were calculated as:

$$AD = (y_i - \hat{y}_i)/n \tag{1}$$

$$MAD = \max\{|y_i - \hat{y}_i|, i = 1,2,\cdots,n\} \tag{2}$$

$$AAD = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

$$RMSE\% = RMSE/\overline{Y} \times 100\% \tag{5}$$

where y_i is the measured value of the response variables (forest canopy closure density and AGB) for plot *i*; $\hat{y}_i$ is the estimated value from the PLS model; $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, and *n* is the number of plots.

The model having the best predicted value (e.g. the minimum RMSE%) was used to estimate the target parameters for the entire study area.

Results

Selecting variables for target parameters based on the bootstrap method

For the canopy closure density estimation model (hereafter referred to as C-Model), 15 of 24 explanatory variables passed the bootstrap test at α=0.05; for the AGB model (hereafter referred to as the B-Model), 18 of 24 explanatory variables passed the bootstrap analyses. Both of the reduced estimation models had 15 common variables, and the variables TM4/2, TM5-7/5+7, and KT2 were only contained in the reduced bootstrap B-Model. The regression coefficients and relevant critical values or thresholds of these remaining variables in both reduced bootstrap models are presented in Fig. 2, respectively.

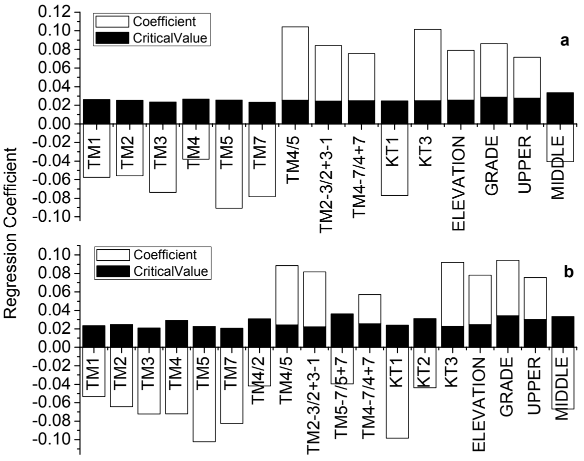

Fig. 2 Standardized regression coefficients and their critical values or thresholds of explanatory variables passing bootstrap tests at α=0.05 in (a) canopy closure density estimation model, and (b) above-ground tree biomass estimation model.

Selecting variables for target parameters based on the VIP criterion

According to the VIP criterion, only 8 of 24 explanatory variables played important roles in canopy closure density estimation. For the B-Model, there were nine important variables, half of the total number of variables yielded by the bootstrap approach. There were seven common variables, including TM3, TM5, TM7, KT1, KT3, TM2-3/2+3-1 and TM4/5, in both reduced VIP models. The variable TM4-7/4+7 was only included in the reduced C-Model while only the variables TM4 and MIDDLE were included in the reduced B-Model. In addition, the variables selected by the VIP criterion were all accepted by bootstrap tests at α=0.05 (Fig. 2). However, the greenness index (that represents the growth status of vegetation) was excluded from both of the reduced VIP models. Fig. 3 shows the VIP values of all 24 explanatory variables in both the full estimation models.

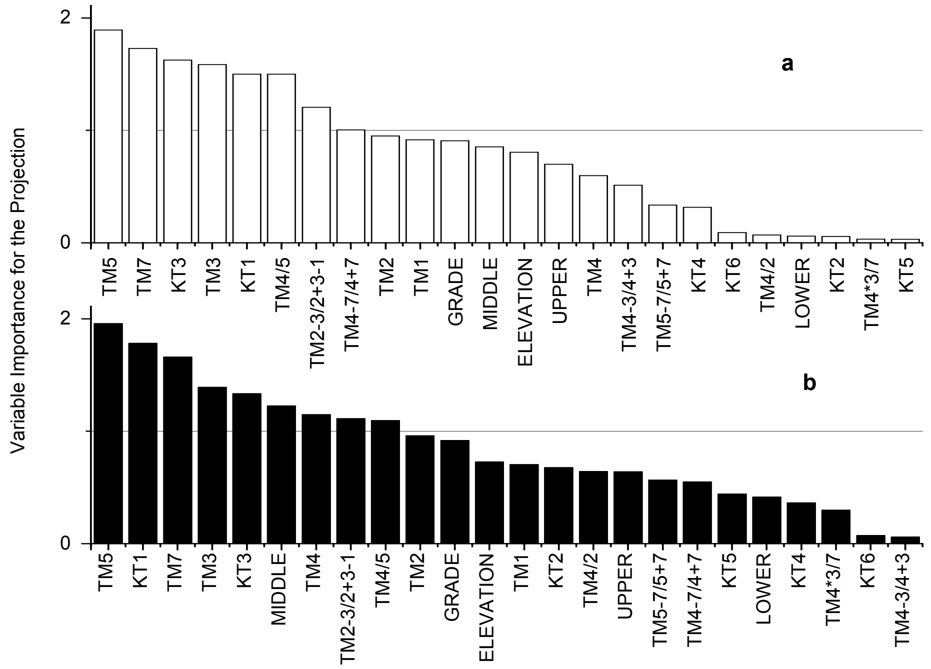

Fig. 3 The VIP values of all 24 explanatory variables in (a) canopy closure density estimate model, and (b) above ground tree biomass model.

Prediction accuracy by validation

Tables 2 and 3 present the prediction accuracy for the full and reduced models with the validation data for two different target variables, canopy closure density and AGB. For the index AD, the predicted values were greater than the actual values for both canopy closure density and AGB. The reduced VIP model showed the best performance among all the C-models (Table 2). But the reduced VIP B-model, regardless of whether the variable MIDDLE was included, was least accurate, yielding 2% lower RMSE% than the others. Canopy closure density was more accurately predicted (~29% RMSE) than was above-ground tree biomass (~57% RMSE). Based on its shorter calculation time, the reduced VIP model was used to estimate AGB. Because exclusion of the terrain variable MIDDLE did not result in less accurate performance of the reduced VIP model (Table 3), the reduced VIP B-Model was referred to as the one without the variable MIDDLE.

Table 2. Prediction performance statistics for three canopy closure density estimate models using validation data (*n*=92)

models	AD	MAD	AAD	RMSE	RMSE%
Full model	0.0337	0.5282	0.1103	0.1484	29.435
Reduced (VIP) model	0.0426	0.5080	0.1107	0.1471	29.191
Reduced (bootstrap) model	0.0331	0.5303	0.1096	0.1478	29.322

Note: AD is average deviation; MAD, AAD, RMSE, and RMSE% signify maximum absolute deviation, average absolute deviation, root mean square error and relative RMSE, respectively.

Table 3. Prediction performance statistics for four above ground tree biomass models using validation data (*n*=92) (t·ha^{-1})

models	AD	MAD	AAD	RMSE	RMSE%
Full model	7.4027	142.93	29.896	39.642	55.752
Reduced (VIP)* model	9.6902	148.09	31.323	41.168	57.899
Reduced (VIP) model	10.072	153.08	31.179	41.606	58.514
Reduced (bootstrap) model	7.3703	143.10	29.483	39.629	55.734

Note: * the variable MIDDLE was excluded from the reduced VIP models. AD is average deviation and MAD, AAD, RMSE, RMSE% signify maximum absolute deviation, average absolute deviation, root mean square error and relative RMSE.

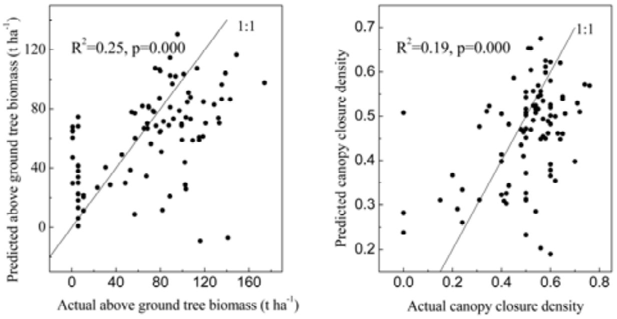

Fig. 4 Scatter plot of predicted versus observed values of canopy closure density (right) and above-ground tree biomass (left) (*n*=92).

Predicted and observed values for canopy closure density and AGB are plotted in Fig. 4, and their agreements were assessed by the determination coefficient. About 19% of the variation in observed canopy closure density was explained by the predictions. About 25% of variation in AGB was explained by the predictions. Both determination coefficients were relatively low, but both were statistically significant ($p < 0.001$).

Application to the entire study area

Using the Landsat pixel digital number covering the study area, forest canopy closure density and AGB were simulated by the reduced VIP models. Fig. 5 shows that forest canopy closure density ranged from 0.4 to 0.7, this was consistent with the field inventory data. The forest was young or middle-age class, thus AGB ranging from 60 to100 t·ha^{-1} was reasonable. Non-forest areas (river, grassland, shrubland) were not excluded before assessment, so these might have caused lower canopy closure density and AGB estimates due to the similarity in satellite data of their digital numbers to those from young forest. Therefore, the average or total AGB and canopy closure density of the entire study area were not considered.

Discussion

Partial least square regression coefficient has complex nonlinear attributes that preclude its testing by widely applied methods such as the t- or F-test (Carreiras et al. 2006). We used the bootstrap method to test partial least square regression parameters due to its non-requirement for data distribution (Serneels and Van Espen 2005; Wang et al. 2006). Our study confirmed that the explanatory variables selected by the bootstrap method were significant for the dependent variables since they ensured high prediction accuracy (Table 2 and Table 3). Furthermore, the reduced VIP models preformed nearly as well as the reduced bootstrap models but had the simplest variable range. This result showed that the VIP criterion applied in this research successfully eliminated insignificant variables and simplified the regression model even though it is a qualitative method in some ways (Wang et al. 2006). Næsset et al. (2005) retained in the model even those variables at VIP≥0.8 for estimating biophysical properties of forest stands, but usage of the lower threshold resulted in inclusion of additional variables and this led to difficulty in model application. Based on the VIP criterion, the explanatory capability of independent variables became explicit. Biotic parameters are best predicted using the middle-infrared bands (Jakubauskas and Price 1997), therefore TM5, TM7 and their combination were included in the reduced model.

We included the atmospherically resistant vegetation index TM2-3/2+3-1 in the reduced model because it is less sensitive to atmospheric effects than is NDVI, and thus is more applicable for vegetated surfaces (Kaufman and Tanre 1992; LeMay et al. 2008). TM4-3/4+3 was, however, excluded from the reduced model. Most terrain factors were also excluded from the model because, in areas with gentle terrain, topographic effects were not major contributors to spectral variation (Jakubauskas and Price 1997). Explanation of 19% of the variation observed in canopy closure density by the predicted value in this study was lower than the 64% to 89% reported by Joshi et al. (2006). Certainly, a better agreement between observed and predicted canopy closure density was also gained by Carreiras et al. (2006). Our lower levels of agreement might have been caused in part by the extreme local variation in mixed forests. In single-species forests such errors possibly play a smaller role as documented by Carreiras et al. (2006). Because the methods used to collect ground-based data can influence predictive models (Fiala et al. 2006), this might also be a source of error.

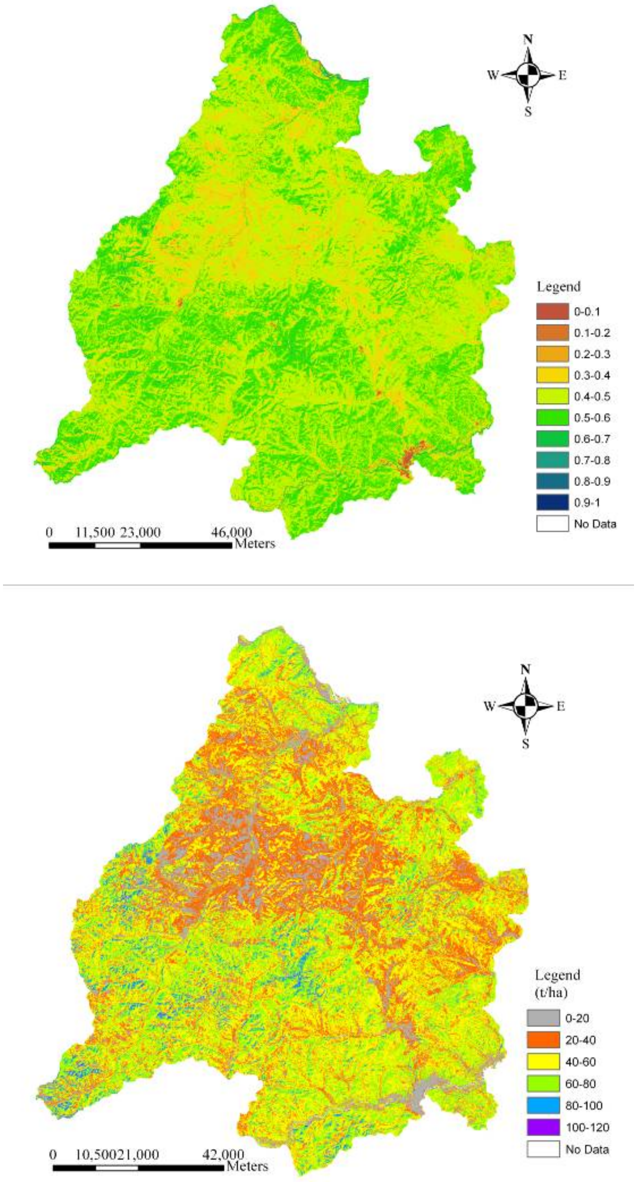

Fig. 5 Estimated canopy closure density (upper) and above-ground biomass in t·ha^{-1} (lower) for each pixel covering study area with non-forest area included

Hemispherical photographs recently produced by fisheye lenses have been used as an alternative in studies of canopy structure (Joshi et al. 2006) and can improve the estimation accuracy of canopy closure density in the field. Similarly, better predicted accuracy of AGB was obtained in pure forests as opposed to mixed forests (Zheng et al. 2004).

The reduced VIP models were then applied to the entire study area and yielded better agreement between observed and predicted data. Though average and total AGB and canopy closure density were not counted, their predicted deviations for forest stands were not greater than the results obtained from validation data: Their estimates tended to be more accurate in predicting mean values while overestimating lower values and underestimating higher values (Reese et al. 2002). To a certain extent, forest structural parameters are related closely to terrain factors due to their determination to redistribution of water, heat, and light (Zhao and Li 2001). Our work showed that, in an area with relative gentle terrain, the terrain factors cannot always lead to better model performance (Table 3).

In our study, the digital number values of remote sensing images were related to field investigation data. However, digital numbers are time- and sensor-specific; therefore, reflectance or radiance values at satellite level due to their insensitivity to temporal and spatial difference may be examined for more perfect relationships between canopy closure density, AGB, and remote sensing data (Jakubauskas and Price 1997; Bruce K. Wylie, personal communion, April 2, 2008). Classifying the images in terms of land use types can improve the reliability of forest parameters.

Acknowledgement
The authors gratefully thank the anonymous reviewers for their helpful comments and suggestions.

References

Carreiras JMB, Pereira JMC, Pereira JS. 2006. Estimation of tree canopy cover in evergreen oak woodlands using remote sensing. *Forest Ecology and Management,* **223**: 45–53.

Cho M, Skidmore A., CorsI F, Van Wieren S, Sobhan I. 2007. Estimation of green grass/herb biomass from airborne hyperspectral imagery using spectral indices and partial least squares regression. *International Journal of Applied Earth Observations and Geoinformation,* **9**: 414–424.

Cohen WB, Maiersperger TK, Gower ST, Turner DP. 2003. An improved strategy for regression of biophysical variables and Landsat ETM+ data. *Remote Sensing of Environment*, **84**: 561−571.

Crist E, Cicone R. 1984. A physically-based transformation of thematic mapper data-the TM tasseled cap. *IEEE Transactions on Geoscience and Remote Sensing,* **22**: 256–263.

Fang JY, Liu GH, Xu SL. 1996. Biomass and net production of forest vegetation in China. *Acta Ecologica Sinica*, **16**: 497−508. (In Chinese with English abstract).

Farifteh J, Van Der Meer F, Atzberger C, Carranza EJM. 2007. Quantitative analysis of salt-affected soil reflectance spectra: A comparison of two adaptive methods (PLSR and ANN). *Remote Sensing of Environment*, **110**: 59–78.

Fiala ACS, Garman SL, GRAY AN. 2006. Comparison of five canopy cover estimation techniques in the western Oregon Cascades. *Forest Ecology and Management,* **232**: 188−197.

Hansen PM, Schjoerring JK. 2003. Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. *Remote Sensing of Environment*, **86**: 542–553.

Hernandez-stefanoni J, Ponce-hernandez R. 2006. Mapping the spatial variability of plant diversity in a tropical forest: comparison of spatial interpolation methods. *Environmental Monitoring and Assessment*, **117**: 307–334.

Jakubauskas ME, PRICE KP. 1997. Empirical relationships between structural and spectral factors of Yellowstone lodgepole pine forests. *Photogrammetric engineering and remote sensing*, **63**: 1375−1381.

Joshi C, De LeeuwJ, Skidmore AK, Van Duren IC, Van Oosten H. 2006. Remotely sensed estimation of forest canopy density: A comparison of the performance of four methods. *International Journal of Applied Earth Observation and Geoinformation*, **8**: 84–95.

Lemay V, Maedel J, Coops NC. 2008. Estimating stand structural details using nearest neighbor analyses to link ground data, forest cover maps, and Landsat imagery. *Remote Sensing of Environment*, **112**: 2578–2591.

Næsset E, Bollandsas OM, Gobakken T. 2005. Comparing regression methods in estimation of biophysical properties of forest stands from two different inventories using laser scanner data. *Remote Sensing of Environment*, **94**: 541–553.

Price K, Guo X, Stiles J. 2002. Optimal Landsat TM band combinations and vegetation indices for discrimination of six grassland types in eastern Kansas. *International Journal of Remote Sensing*, **23**: 5031−5042.

Reese H, Nilsson M, Sandström P, Olsson H. 2002. Applications using estimates of forest parameters derived from satellite and forest inventory data. *Computers and Electronics in Agriculture*, **37**: 37–55.

Serneels S, Van Espen P. 2005. Bootstrap confidence intervals for trilinear partial least squares regression. *Analytica Chimica Acta*, **544**: 153−158.

Wang HW, Wu ZB, Meng J (Eds.). 2006. *Partial Least-Squares Regression – Linear and Nonlinear Methods*. Beijing: National Defense Industry Press. pp.35−123 (In Chinese).

WU FZ (Ed.). 1996. *Forest Mensuration*. Beijing: Chinese Forestry Publishing House, pp.174–202 (In Chinese).

Wulder M, Han T, White J, Sweda T, Tsuzuki H. 2007. Integrating profiling LIDAR with Landsat data for regional boreal forest canopy attribute estimation and change characterization. *Remote Sensing of Environment*, **110**: 123−137.

Zhao XW, Li CG (Eds.). 2001. *Quantitative Estimation of Forest Resource Based on "3S": Theory, Method Application and Software*. Beijing: China Science and Technology Press, pp. 12–46 (In Chinese).

Zheng D, Rademacher J, Chen J, Crow T, Bresee M, Le Moine J, Ryu SR. 2004. Estimating aboveground biomass using Landsat 7 ETM+ data across a managed landscape in northern Wisconsin, USA. *Remote Sensing of Environment*, **93**: 402–411.